KELOMPOK:

Moh. Naufal Faqih 10222044

Firman Firdaus 10222033

Ryan Azis S. 10222041

Studi Kasus: Analisis Data Pasien Klinik Kesehatan

1. Memuat Dataset Pasien

```
df = pd.read_csv('DATASET\data_pasien_diabetes.csv')
df.head()
```

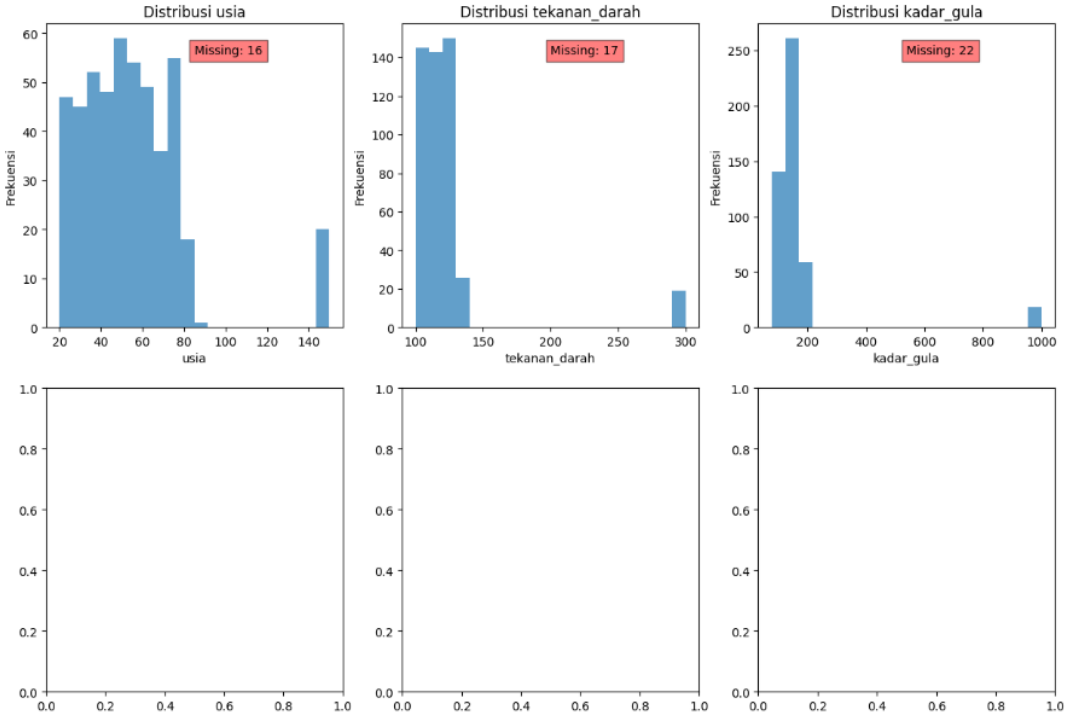
	id_pasien	nama	usia	tekanan_darah	kadar_gula	jenis_kelamin	alamat	diagnosa
0	P0001	Anto Sartika	62.0	123.0	127.0	laki-laki	Jl. Melur No.71, Medan	Pra-diabetes
1	P0002	Nina Wati	70.0	121.0	169.0	Laki	Jalan Kenanga No.26, Medan	Diabetes Tipe 2
2	P0003	Rahmat Aminah	38.0	100.0	161.0	Laki	Jl. Melur No.47, Surabaya	Pra-diabetes
3	P0004	Siti Saputra	25.0	100.0	166.0	L	Jl. Anggrek No.52, Medan	Tidak Diabetes
4	P0005	Andi Wati	49.0	110.0	131.0	perempuan	Perum Griya Indah No.74, Medan	Diabetes Tipe 2

2. Mendeteksi Missing Value

```
print("Jumlah Missing Value per kolom:")
print(df.isnull().sum())
```

```
Jumlah Missing Value per kolom:
id_pasien         0
nama              0
usia             16
tekanan_darah    17
kadar_gula       22
jenis_kelamin     0
alamat           45
diagnosa          0
dtype: int64
```

3. Visualisasi Sebelum Data Dibersihkan



4. Menangani beberapa Missing Value

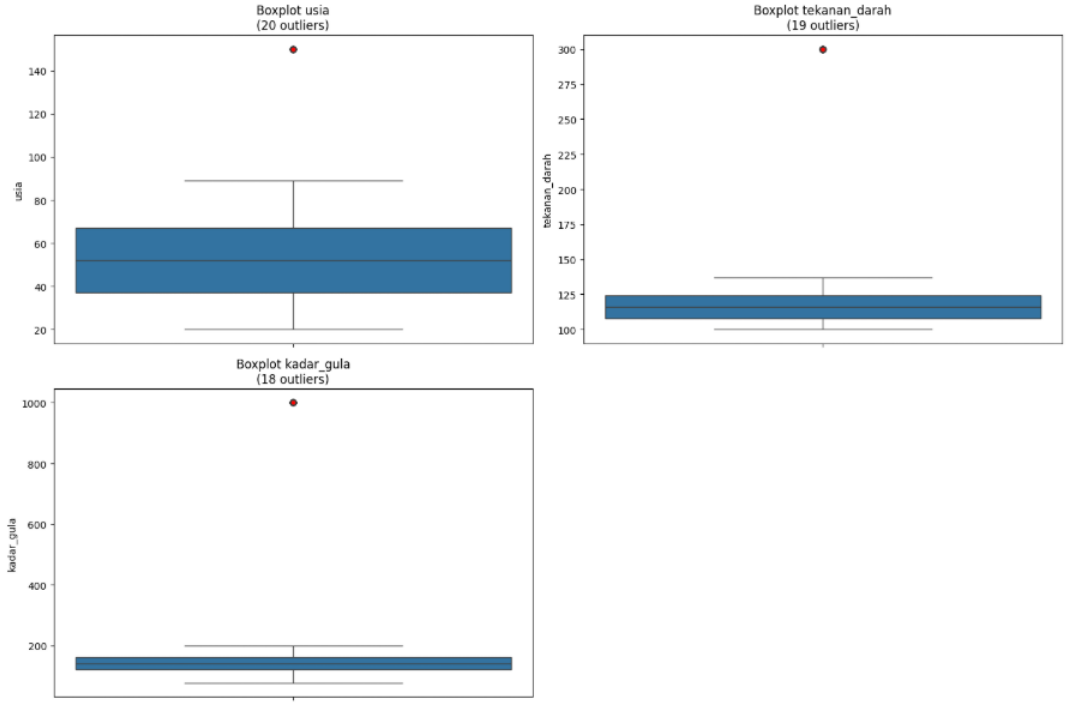
```
df['usia'] = df['usia'].fillna(df['usia'].mean())
df['kadar_gula'] = df['kadar_gula'].fillna(df['kadar_gula'].median())
df['alamat'] = df['alamat'].fillna(df['alamat'].mode()[0])
df['tekanan_darah'] = df['tekanan_darah'].fillna(df['tekanan_darah'].median())
print("\nSetelah diisi dengan rata-rata:")
print(df.isnull().sum())
```

```
Setelah diisi dengan rata-rata:
id_pasien        0
nama             0
usia             0
tekanan_darah    0
kadar_gula       0
jenis_kelamin    0
alamat           0
diagnosa         0
dtype: int64
```

	id_pasien	nama	usia	tekanan_darah	kadar_gula	jenis_kelamin	alamat	diagnosa
0	P0001	Anto Sartika	62.000000	123.0	127.0	laki-laki	Jl. Melur No.71, Medan	Pra-diabetes
1	P0002	Nina Wati	70.000000	121.0	169.0	Laki	Jalan Kenanga No.26, Medan	Diabetes Tipe 2
2	P0003	Rahmat Aminah	38.000000	100.0	161.0	Laki	Jl. Melur No.47, Surabaya	Pra-diabetes
3	P0004	Siti Saputra	25.000000	100.0	166.0	L	Jl. Anggrek No.52, Medan	Tidak Diabetes
4	P0005	Andi Wati	49.000000	110.0	131.0	perempuan	Perum Griya Indah No.74, Medan	Diabetes Tipe 2
5	P0006	Linda Prasetyo	40.000000	122.0	137.0	laki-laki	Jl. Anggrek No.3, Bandung	Tidak Diabetes
6	P0007	Rudi Agustin	57.000000	114.0	125.0	perempuan	Jl. Melur No.15, Bandung	Tidak Diabetes
7	P0008	Rahmat Wati	69.000000	120.0	150.0	L	Jl. Melur No.59, Jakarta	Diabetes Tipe 2
8	P0009	Rudi Aminah	54.442149	108.0	149.0	L	Jalan Mawar No.40, Bandung	Diabetes Tipe 2
9	P0010	Siti Saputra	71.000000	125.0	120.0	p	Jl. Merdeka No.19, Yogyakarta	Pra-diabetes

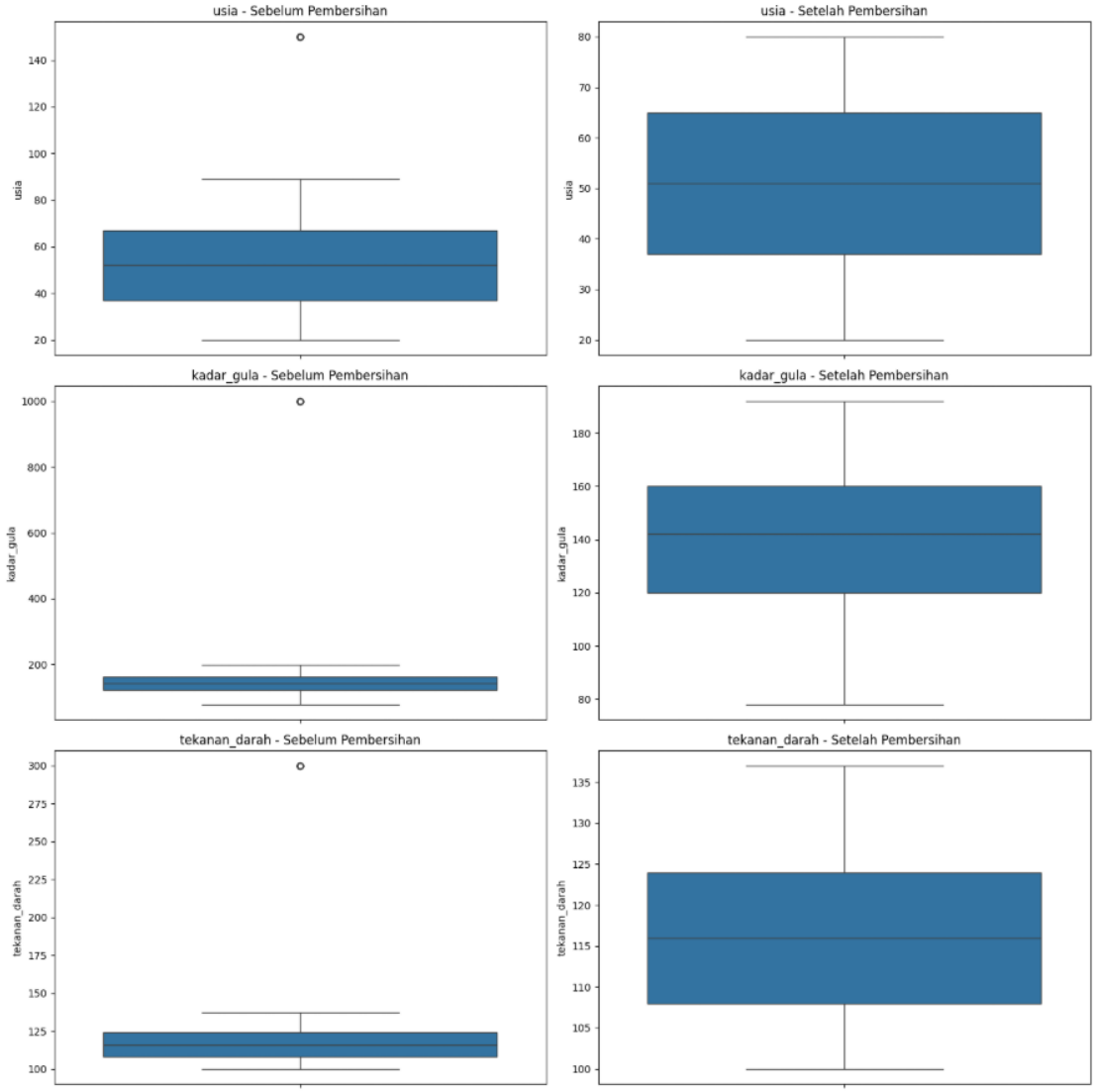
5. Mendeteksi Outlier

```
Jumlah outlier per kolom (metode Z-score):
usia: 20 outlier
tekanan_darah: 19 outlier
kadar_gula: 18 outlier
```



6. Tangani Outlier dengan Teknik Remove

```
Jumlah outlier yang ditemukan per kolom:
  usia: 20 outlier
  kadar_gula: 18 outlier
  tekanan_darah: 19 outlier
Total baris yang dihapus: 51
Jumlah baris sebelum pembersihan: 500
Jumlah baris setelah pembersihan: 448

Statistik sebelum pembersihan:
             usia   kadar_gula  tekanan_darah
count  500.000000   500.000000     500.000000
mean    54.442149   170.870000     122.574000
std     25.749424   162.189865      36.386157
min     20.000000    78.000000     100.000000
25%     37.000000   121.000000     108.000000
50%     52.000000   142.000000     116.000000
75%     67.000000   163.000000     124.000000
max    150.000000  1000.000000     300.000000

Statistik setelah pembersihan:
             usia  kadar_gula  tekanan_darah
count  448.000000  448.000000     448.000000
mean    50.568376  139.685268     115.517857
std     17.036562   24.402416       9.026416
min     20.000000   78.000000     100.000000
25%     37.000000  120.000000     108.000000
50%     51.000000  142.000000     116.000000
75%     65.000000  160.250000     124.000000
max     80.000000  192.000000     137.000000
```

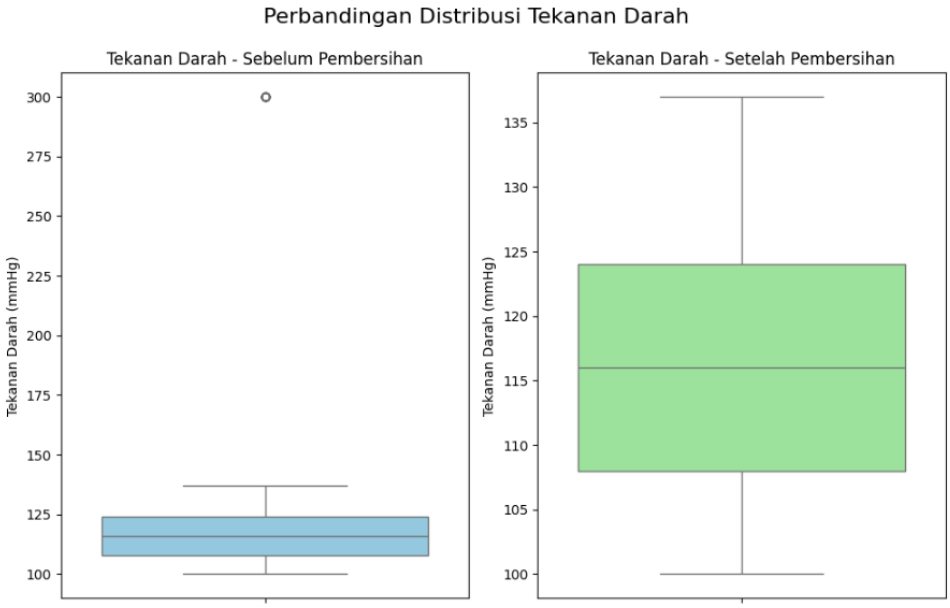


	id_pasien	nama	usia	tekanan_darah	kadar_gula	jenis_kelamin	alamat	diagnosa
0	P0001	Anto Sartika	62.000000	123.0	127.0	laki-laki	Jl. Melur No.71, Medan	Pra-diabetes
1	P0002	Nina Wati	70.000000	121.0	169.0	Laki	Jalan Kenanga No.26, Medan	Diabetes Tipe 2
2	P0003	Rahmat Aminah	38.000000	100.0	161.0	Laki	Jl. Melur No.47, Surabaya	Pra-diabetes
3	P0004	Siti Saputra	25.000000	100.0	166.0	L	Jl. Anggrek No.52, Medan	Tidak Diabetes
4	P0005	Andi Wati	49.000000	110.0	131.0	perempuan	Perum Griya Indah No.74, Medan	Diabetes Tipe 2
5	P0006	Linda Prasetyo	40.000000	122.0	137.0	laki-laki	Jl. Anggrek No.3, Bandung	Tidak Diabetes
6	P0007	Rudi Agustin	57.000000	114.0	125.0	perempuan	Jl. Melur No.15, Bandung	Tidak Diabetes
7	P0008	Rahmat Wati	69.000000	120.0	150.0	L	Jl. Melur No.59, Jakarta	Diabetes Tipe 2
8	P0009	Rudi Aminah	54.442149	108.0	149.0	L	Jalan Mawar No.40, Bandung	Diabetes Tipe 2
9	P0010	Siti Saputra	71.000000	125.0	120.0	p	Jl. Merdeka No.19, Yogyakarta	Pra-diabetes

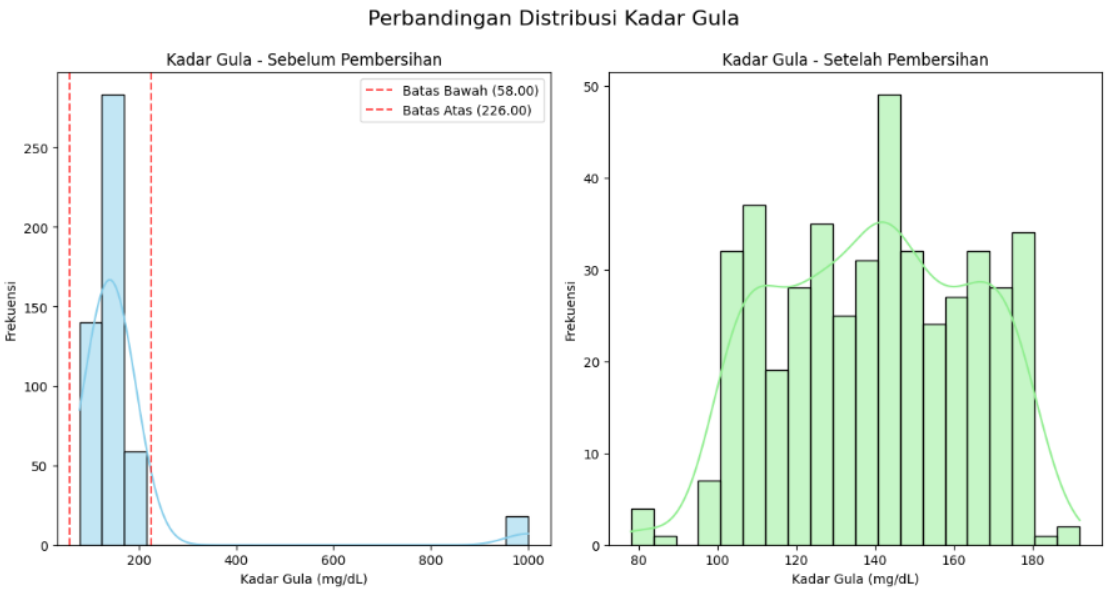
7. Normalisasi Kolom Gender dan Diagnosa

```
Nilai unik jenis_kelamin sebelum normalisasi: ['laki-laki' 'Laki' 'L' 'perempuan' 'p' 'P']
Nilai unik diagnosa sebelum normalisasi: ['Pra-diabetes' 'Diabetes Tipe 2' 'Tidak Diabetes' 'Diabetes Tipe 1']

Nilai unik jenis_kelamin setelah normalisasi: ['L' 'P']
Nilai unik diagnosa setelah normalisasi: ['Pra-diabetes' 'Diabetes Tipe 2' 'Normal' 'Diabetes Tipe 1']

Jumlah per kategori jenis kelamin:
jenis_kelamin
P    254
L    194
Name: count, dtype: int64

Jumlah per kategori diagnosa:
diagnosa
Normal             117
Diabetes Tipe 2    112
Diabetes Tipe 1    111
Pra-diabetes       108
Name: count, dtype: int64
```

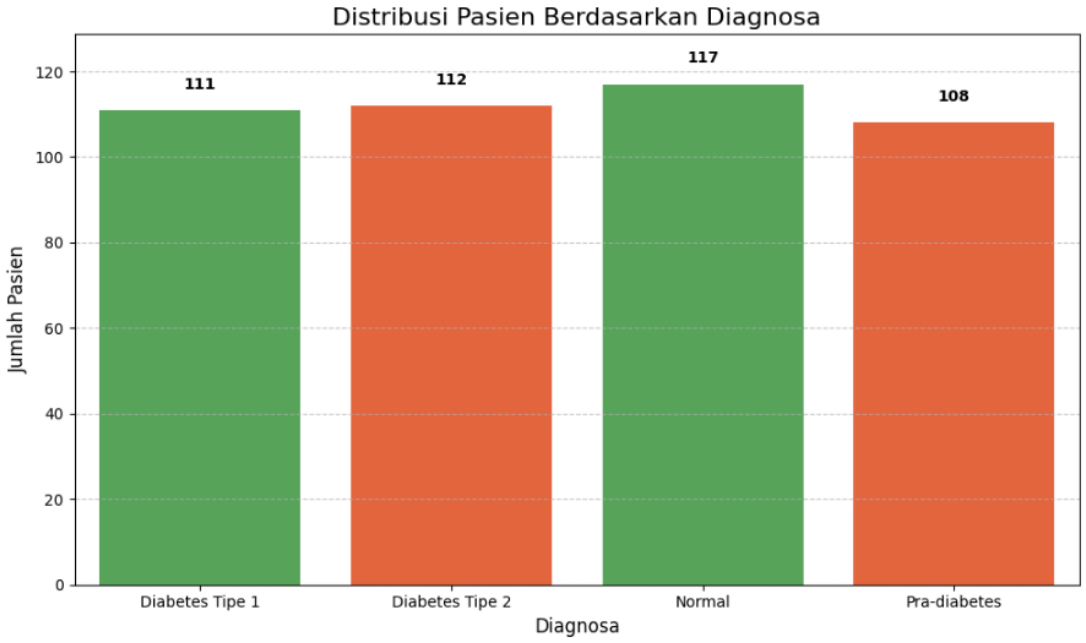
8. Visualisasi




```
Statistik tekanan_darah sebelum pembersihan:
count    500.000000
mean     122.574000
std       36.386157
min      100.000000
25%      108.000000
50%      116.000000
75%      124.000000
max      300.000000
Name: tekanan_darah, dtype: float64

Statistik tekanan_darah setelah pembersihan:
count    448.000000
mean     115.517857
std        9.026416
min      100.000000
25%      108.000000
50%      116.000000
75%      124.000000
max      137.000000
Name: tekanan_darah, dtype: float64

Jumlah outlier pada tekanan_darah: 0
Batas bawah: 84.0, Batas atas: 148.0
```




```
Statistik kadar_gula sebelum pembersihan:
count     500.000000
mean      170.870000
std       162.189865
min        78.000000
25%       121.000000
50%       142.000000
75%       163.000000
max      1000.000000
Name: kadar_gula, dtype: float64

Statistik kadar_gula setelah pembersihan:
count    448.000000
mean     139.685268
std       24.402416
min       78.000000
25%      120.000000
50%      142.000000
75%      160.250000
max      192.000000
Name: kadar_gula, dtype: float64

Jumlah outlier pada kadar_gula: 0
Batas bawah: 58.00, Batas atas: 226.00
```




```
Persentase pasien per kategori diagnosa:
Diabetes Tipe 1: 24.8%
Diabetes Tipe 2: 25.0%
Normal: 26.1%
Pra-diabetes: 24.1%
```

9. Kesimpulan
    a. Penanganan Missing Value

       Pengisian missing values dengan nilai statistik yang tepat (mean, median, mode) memastikan:

        - Dataset lengkap tanpa nilai kosong, sehingga analisis dapat dilakukan tanpa gangguan.
        - Lebih representatif karena menggunakan nilai-nilai yang mendekati distribusi data asli.
        - Jumlah data tetap terjaga mempertahankan kekuatan statistik dataset.

    b. Penannganan Outlier

       Penghapusan outlier memberikan dampak positif pada kualitas analisis:

        - **Distribusi Data Lebih Normal**: Terlihat dari visualiasi boxplot dan histogram yang menunjukkan distribusi lebih seimbang setelah pembersihan.
        - **Statistik Deskriptif Lebih Representatif**: Nilai mean dan standar deviasi tidak lagi terpengaruh oleh data yang ekstrem nilainya.
        - **Peningkatan Keandalan Analisis**: Mengurangi risiko bias pada proses analisis selanjutnya.

    c. Perubahan Pada Visualiasi

       Perbandingan visualisasi menunjukkan perubahan signifikan

        - **Boxplot Tekanan Darah**: Hilangnya titik-titik ekstrem pada visualisasi setelah pembersihan, menunjukkan distribusi yang lebih bagus.
        - **Histogram Kadar Gula**: Kurva distribusi menjadi lebih mendekati distribusi normal setelah penghapusan outlier.
        - **Bar Chart Diagnosa**: Menunjukkan perbandingan proporsi pasien dengan diabetes dan non-diabetes yang lebih akurat.

    d. Normalisasi Data Kategorikal

       Standarisasi nilai pada kolom jenis_kelamin dan diagnosa menghasilkan:

        - **Konsistensi Format Data**: Semua nilai gender menjadi format standar ‘L’ dan ‘P’.
        - **Kemudahan Interpretasi**: Nilai diagnosa yang konsisten memudahkan analisis dan visualisasi.

- **Peningkatan Keakuratan Perhitungan**: Pengelompokkan yang tepat saat analisis kategorikal.

Secara keseluruhan, proses pembersihan data telah mengubah dataset yang semula memiliki missing values dan outlier menjadi dataset yang lebih berkualitas dan representative. Visualisasi setelah pembersihan menunjukkan gambaran yang lebih akurat tentang distribusi dan karakteristik pasien diabetes yang sangat penting untuk analisis medis dan pengambilan keputusan klinis.